1998 Washington State Population Survey

TECHNICAL REPORT #2 – Weighting Procedure

Office of Financial Management
Forecasting

JANUARY 7, 1999

THIS REPORT is one of several technical appendices prepared by the Office of Financial Management to supplement the 1998 Washington State Population Survey (SPS). The survey was funded by the 1997 Legislature to provide social, demographic, and economic information about Washington residents that would otherwise not be available between the 1990 and 2000 federal census years. A background report on the SPS titled, *1998 Washington State Population Survey Data Report* (September 21, 1998), is available electronically at http://www.wa.gov/ofm/ under Population/Data, or by calling OFM at (360) 902-0599.

Responses to the survey were obtained from telephone interviews of 7,279 households which represent the state population as a whole. The survey was designed by OFM and conducted by the Washington State University Social and Economic Sciences Research Center (SESRC). Telephone interviews were conducted in Spring 1998.

The Weighting Procedure

Sample surveys, especially telephone sample surveys, usually contain non-response errors which are caused by under-coverage or refusal to participate. If the under-coverage and refusal exist systematically in certain segments of the population, the data may be biased. One way to partially reduce this bias is to use a weighting procedure to cause the survey estimates to match the known characteristics of the population. This weighting procedure is commonly known as post-stratification.

A post-stratification procedure usually involves creating two classification tables, one for the survey data and one for the data that are used as controls. Both tables must have the same number of cells with each cell representing a unique combination of the attributes of the control characteristics (or variables). An inverse ratio is then taken of the counts in corresponding cells in the survey data table and the control table. The result of this estimation is then recorded in a weight variable which can be applied in most statistical analyses to the survey data to generate inferences about the population.

This post-stratification procedure was used in the SPS weighting. Data from the *1998 County Population Estimates by Race/Ethnicity, Age, and Sex* produced by OFM were used as the controls. The following discussion provides a detailed description of this procedure as adopted for the SPS.

Selecting Control Variables

The first step in the SPS post-stratification was to identify which variables would be used as the control variables. Generally, key demographic variables would be used as the control variables in post-stratification. Depending on the survey and sample designs, other variables may be introduced as well. However, more variables and larger numbers of values of those variables would result in smaller cell sizes and sometimes in empty cells. When this happens, matching can be very difficult. Therefore, the number of control variables has to be kept reasonably small.

The post-stratification of the SPS data was accomplished in two phases. The first phase involved simultaneous stratification of five control variables. These are region, sample type, age, sex, and race/ethnicity. The second phase involved stratifying the data by controlling for household size. Household size was stratified independent of the other five variables because household size information was not available in the control data.

Both the region and sample type variables were selected because of the sample design. Age, sex, and race/ethnicity were selected because they are basic demographic variables. Household size was included because single-person households are generally under-represented in RDD surveys.

The control variables and their values are as follows:

1. Region

(1)	Region 1:	Island, San Juan, Skagit, Whatcom
(2)	Region 2:	Clallam, Cowlitz, Grays Harbor, Jefferson, Klickitat, Lewis, Mason, Pacific, Skamania, Wahkiakum
(3)	Region 3:	King
(4)	Region 4:	Kitsap, Pierce, Snohomish, Thurston
(5)	Region 5:	Clark
(6)	Region 6:	Adams, Asotin, Chelan, Columbia, Douglas, Ferry, Garfield, Grant, Kittitas, Lincoln, Okanogan, Pend Oreille, Stevens, Walla Walla, Whitman
(7)	Region 7:	Spokane
(8)	Region 8:	Benton, Franklin, Yakima

2. Sample Type

(1) General Population Sample (main sample)
(2) Expanded Sample of Racial Minorities

3. Age

(1) 0-9 years
(2) 10-19 years
(3) 20-29 years
(4) 30-39 years
(5) 40-49 years
(6) 50-59 years
(7) 60-69 years
(8) 70-79 years

(9) 80 years or older

4. Sex

(1) Male
(2) Female

5. Race/Ethnicity

Hispanic
(1) Hispanic

Non-Hispanics
(1) Black
(2) Native American
(3) Asian and Pacific Islander
(4) White

6. Household Size

(1) 1 person
(2) 2 persons
(3) 3 persons
(4) 4 persons
(5) 5 persons
(6) 6 persons
(7) 7 persons
(8) 8 persons
(9) 9 persons
(10) 10 or more persons

Data Imputation of the Control Variables

Post-stratification of sample survey data requires that control variables contain no missing data. A case is said to contain missing data if the response to one or more questions is "refusal," "don't know," or "not ascertained" due to questionnaire design defects, interviewer errors, or data processing errors. If a case contains missing data in any of the control variables, it would result in being rejected by the computation process. To ensure control variables have no missing data, researchers commonly conduct data imputation for the variables that do contain missing values. This practice was adopted in the SPS post-stratification.

A frequency review indicated that three of the six selected control variables (sample type, region, and household size) in SPS post-stratification contained no missing data. The other three (age, sex, and race/ethnicity), however, did contain missing data to varying degrees.

Two main imputation techniques were employed with these three variables. One was deductive imputation and the other was simple random imputation. A deductive imputation involves deducing a response by checking the logic or context of questions before and/or after the question in which a response is missing. For example, if the response on sex is missing for a household member, but the relationship question indicates that this person is the daughter of the respondent, then the sex is set to be female.

The simple random imputation method assigns a response at random within a specified range. This usually involves the use of a random number function.

All three demographic variables underwent a combination of the two imputation methods. For the age variable, the respondent's age was imputed first. Other household members' age and relationship to the respondent were checked to see if an age could be deduced for the respondent. If that attempt failed, then a random age was selected from the range of 18 to 75.

The age variable was then imputed for other household members. Similarly, the relationship of the other members to the respondent was checked to deduce a response. If the relationship check failed, then the member's marital status, education attainment, and military service status were checked to determine whether this individual was a child or an adult. If it was determined that the member was an adult, then a random age from 18 to 75 was assigned. If the member was determined to be a child, a random age from 0 to 17 was assigned. If this attempt failed, too, then a random age was selected from the range of 0 to 75.

The sex variable, as mentioned earlier, was checked for a member's relationship to the respondent. If this attempt resulted in no response, then a random choice was assigned between male and female.

The race/ethnicity variable is a composite variable constructed from two original variables: race and Hispanic origin. The imputation was performed on the two original variables and was carried out in two stages. In the survey, the question on Hispanic origin was asked first and then followed by the question on race. The race question contained five response categories: Black, Native American, Asian or Pacific Islander, White, and other. The proportion choosing the "other" category was quite large, 12.3 percent of all person records. Many of these were persons of Hispanic origin. Some were of mixed races. There were still some who were of one of the four listed races but chose to answer it differently. For example, instead of choosing "White," a respondent might say "Dutch." Rather than coding it, some interviewers recorded the response as given. In addition to the "other" category in the race and Hispanic variables that needed to be imputed, the refusal, "don't know," and not-ascertained cases were also imputed. These cases constituted 1 percent of the total person records.

The "other" category was first used to determine a person's race and Hispanic origin if "other" was the response. Race and Hispanic origin were determined simultaneously. If any description of Hispanic origin was found in the response, that person was classified as Hispanic. Race was determined by the following rules:

- If only one race description was identified, that description was coded accordingly. If this one description was Hispanic, then it was coded as "White."
- If more than one description (excluding Hispanic) was identified, then the following rules were applied:
    1. If the response involved one minority race and White, then the respondent was coded as a minority race.

2. If the response involved more than one minority race, then the first minority race that appeared in the description was used.

The remaining missing cases were imputed by a technique called proportional random imputation. It was assumed that these remaining cases would have a distribution pattern similar to that of newly identified cases in the "other" category. In the proportional random imputation, cases that contained missing race and Hispanic origin information were assigned a random number ranging from 0 to 1. Then the proportions of races and persons of Hispanic origin as found in the "other" category were applied to this random number to determine the proportion of each race and people of Hispanic origin in the remaining missing cases.

The item non-response rates for the three variables were small. For age, the non-response rate was 2.5 percent. The rate was only 0.2 percent for the sex variable. The race variable had only 1 percent non-response. However, because the "other" category had to be imputed (or recoded), the final imputation rate turned out to be 13.4 percent (See Table 1). Similarly, the Hispanic origin variable had a non-response rate of 0.5 percent, but due to the imputation of the "other" category in the race question, the final imputation rate was 7.5 percent.

TABLE 1
Items Non-response Rates

Variable	Non-response Rate (percent)	Imputation Rate (percent)
Age	2.5	2.5
Sex	0.2	0.2
Race	1.0	13.4
Hispanic Origin	0.5	7.5

Population Controls

The *1998 County Population Estimates by Race/Ethnicity, Age, and Sex* prepared by OFM was used to create the population controls. The OFM county population estimates contained aggregates for each five-year age group by sex. Each of the age-sex groupings is crossed with four race categories for non-Hispanics and four for Hispanics. The racial categories are: White, Black, Native American, and Asian/Pacific Islander.

Aggregating the County Population Estimates

The original age groupings in *County Population Estimates* were collapsed into 10-year intervals. A total of nine age groupings were created as a result: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80 and over. The 39 counties were grouped into eight regions based on the SPS sample stratification specifications (see *Selecting Control Variables* in this document). All of the four Hispanic racial groups were combined into one and were referred to as Hispanic, so the new race/ethnic classification scheme contained five categories: non-Hispanic White, non-Hispanic Black, non-Hispanic Native American, non-Hispanic Asian or Pacific Islander, and Hispanic.

Constructing the Indicator of Residence in an Area with a High Minority Concentration

The Summary Tape File 1A (STF1A) of the 1990 Census for Washington State was used to obtain proportions of racial/ethnic minorities living in census tracts of high concentration of such

populations. A high concentration tract is defined as a tract in which a particular racial/ethnic minority population constituted 40 percent or more of the total population in that tract. This criterion was used by SESRC to draw the expanded minority sample. For each county, the proportions of minority populations in high concentration tracts and in non-high concentration tracts were recorded separately. These proportions were then applied to the data in *1998 County Population Estimates* to obtain separate estimates of minorities in high-concentration area and in non-high-concentration areas. Regional proportions were then obtained by aggregating the county estimates within a region.

Preparing Survey Data Cells

A set of 20 variables was first created to flag all possible combinations of sex (2), race (5), and sample (2). Also, an age variable (9) was created which has the same categories as those in the population controls. A region variable (8) was then created with the county location information based on the sample stratification specification. These five variables contributed to a total of 1,440 cells of unique combinations. A weighting identification number was then assigned to individual records based on the unique combinations.

Data were then sorted by region and age category. For each age category within a region, a frequency was counted for each of the sex-race-sample combination. These frequencies were output into a separate file. The final output contains the same components as the output from the population controls.

Combining Population Controls and Survey Data Cells and Assigning Weights

Merging the Output Files

The output files from the population controls and the survey data were then merged resulting in two sets of 1,440 cells, one for the population controls and one for the survey data.

Treating Empty Cells

Because of the limited sample size, some of the cells for the survey data were empty. Most of these empty cells were due to the sample-type variable. The following procedures were used to treat the empty cells:

a. If a cell of the survey data was empty in the expanded sample but the corresponding cell in the general population sample was not empty, then its counterpart (high-concentration) in the population controls was set to zero. The original value in the counterpart was then added to the corresponding non-high-concentration cell.

b. If a cell of the survey data was found empty in the general population sample but not in the corresponding expanded sample cell, then its counterpart (non-high-concentration) in the population controls was set to zero. The original value in the counterpart was then added to the corresponding high-concentration cell.

c. If both the general population sample cell and the expanded sample cell in the survey data were found to be empty, then both the high-concentration cell and the non-high-concentration cell in the population controls were set to zero. However, the original values in the population controls cells were recorded and redistributed into other age categories of the same sex-race categories.

Assigning Weights

A weight was then assigned to the non-empty cells in the survey data by taking the inverse ratio of the survey counts and the corresponding population controls counts. The empty cells in the survey data were assigned zero weight. All persons in the same cell were thus assigned the same weight.

Adjusting the Weight for Household Size

The weight was then adjusted for the household size. The adjustment was only made at the state level because of limitations in data availability. The Washington State household size data from the March *Current Population Survey of 1996 and 1997* (CPS) by the U.S. Bureau of the Census were used as controls. The rationale for adjusting for household size was that non-response rates tend to be higher in single-person households than in households with two or more persons.

An average proportion was taken between the two years' CPS data for each household size. These proportions were applied to the weight variable in SPS so that the proportions of different household sizes are reflective of those of the 1996 and 1997 CPS averages. Because the adjustment of household size was made after all other control variables were considered, it slightly changed the race and Hispanic origin distribution from the population control estimates. A further adjustment was made at the state level to align the race and Hispanic origin estimates in the survey with the state population control estimates.

Comparing Population Characteristics Before and After Post-stratification

A comparison of the population estimates of a random sample before and after post-stratification will provide an indication of the extent of sample bias. Table 2 contains the estimates from the general population sample on five of the population characteristics for the state and the eight regions both before and after the post-stratification. The following observations can be made from this table:

- Overall, the sample is fairly representative of the state population. The difference between the estimates before and after the post-stratification ranges from 0.2 to 2.8 percentage points for the state and from 0 to 4.5 percentage points for the regions.
- The sex estimates contain the smallest difference for the state. The difference is also relatively small for the regions.
- Children are consistently over-represented in the sample, while single-person households are consistently under-represented. These factors appear to be related. The under-representation of single-person households means an over-representation of households with more than one person. The latter are more much likely to have children present in the household.
- Hispanic and Native American populations are over-represented in the sample. No particular patterns are found for the Black, Asian or Pacific Islander, and White populations in the regional estimates. However, at the state level, the Black and Asian or Pacific Islander populations are under-represented and the White population is over-represented.

TABLE 2

Population Characteristics of the General Population Sample Before and After Post-stratification

(U=proportion before post-stratification W=proportion after post-stratification)

		Age		Sex		Hispanic Origin		Race				Household Size	
		0-18	19+	M	F	Yes	No	Black	Indian	API	White	1 person	2+
State	u	28.9	71.1	49.6	50.4	7.5	92.5	2.1	3.0	3.4	91.5	7.5	92.5
	w	26.6	73.4	49.4	50.6	5.0	95.0	3.1	1.6	5.6	89.7	10.3	89.7
Region 1	u	26.4	73.6	50.2	49.8	5.3	94.7	0.5	2.5	2.5	94.5	8.2	91.8
	w	26.1	73.9	49.6	50.4	5.6	94.4	1.3	1.7	3.3	93.7	10.9	89.1
Region 2	u	28.3	71.7	49.5	50.5	4.3	95.7	0.4	4.4	1.8	93.4	7.2	92.8
	w	25.7	74.3	49.6	50.4	3.3	96.7	0.7	2.8	1.8	94.7	9.5	90.5
Region 3	u	25.2	74.8	49.9	50.1	6.1	93.9	4.7	2.1	7.6	85.6	9.6	90.4
	w	24.4	75.6	49.1	50.9	3.2	96.8	5.0	1.2	9.2	84.6	12.3	87.7
Region 4	u	29.1	70.9	49.4	50.6	6.1	93.9	4.4	2.8	4.4	88.4	6.9	93.1
	w	28.0	72.0	49.8	50.2	3.6	96.4	3.8	1.5	6.2	88.5	8.9	91.1
Region 5	u	28.6	71.4	48.2	51.8	5.0	95.0	2.3	2.5	4.4	90.8	6.7	93.3
	w	28.3	71.7	49.2	50.8	3.8	96.2	1.6	1.0	4.0	93.4	8.2	91.8
Region 6	u	30.7	69.3	51.3	48.7	13.0	87.0	0.9	3.9	3.3	91.9	8.2	91.8
	w	27.9	72.1	49.4	50.6	11.7	88.3	0.8	2.0	2.4	94.8	11.3	88.7
Region 7	u	30.9	69.1	48.3	51.7	3.6	96.4	1.7	2.4	1.6	94.3	7.4	92.6
	w	26.4	73.6	49.0	51.0	2.4	97.6	1.5	1.7	2.5	94.3	10.9	89.1
Region 8	u	31.1	68.9	50.1	49.9	17.8	82.2	1.6	3.2	2.1	93.1	5.8	94.2
	w	28.4	71.6	48.9	51.1	17.7	82.3	1.9	2.2	2.8	93.1	8.5	91.5

Table 3 contains the comparison of the estimates before and after post-stratification for the expanded sample. At the state level, Table 3 shows larger differences between the two groups of estimates than Table 2. Here are the general observations at the state level from Table 3:

- Children are over-represented in the survey sample, but so are single-person households.
- Females are over-represented than males.
- Race and Hispanic origin estimates understandably show a much larger disparity because the design intentionally over-sampled the racial/ethnic minority groups.

TABLE 3

Population Characteristics of the Expanded Sample Before and After Post-stratification

(U=proportion before post-stratification W=proportion after post-stratification)

		Age		Sex		Hispanic Origin		Race				Household Size	
		0-18	19+	M	F	Yes	No	Black	Indian	API	White	1 person	2+
State	u	33.6	66.4	49.4	50.6	18.9	81.1	24.9	34.0	20.0	21.1	6.6	93.4
	w	28.0	72.0	53.2	46.8	29.7	70.3	9.7	8.9	13.4	68.0	3.7	96.3